The Security Margin: a measure of Source Distinguishability under Adversarial Conditions

M. Barni, B.Tondi

Dept. of Information Engineering and Mathematics - University of Siena , ITALY, e-mail: barni@dii.unisi.it, benedettatondi@gmail.com

We analyze the distinguishability of two sources under adversarial conditions, when the error exponents of type I and type II error probabilities are allowed to take an arbitrarily small value. By exploiting the parallelism between the attacker's goal and optimal transport theory, we introduce the concept of Security Margin defined as the maximum average per-sample distortion introduced by the attacker for which two sources can be reliably distinguished.

GOAL AND MOTIVATION

Long term goal: measure the **security** of a digital system under adversarial conditions

Specific goal: understand the fundamental limits of source identification (hypothesis testing) in presence of adversary

Practical motivation: define a parameter that measures the security of a digital system against a certain type of adversary (rational and intelligent) and under certain conditions.

THE SOURCE IDENTIFICATION GAME

The Source Identification problem with known sources (SI_{ks})

Defender's (D's) goal: to decide if an observed sequence x^n is generated by X (Hp0) or Y

Attacker's (A's) goal: to modify a given sequence y^n in order to mislead the analyst, subject to a distortion constraint

KNOWN RESULTS

Summary of the analysis and the findings in [1] by exploiting *transportation theory* for A's strategy

Limitations for D:
1) 1st order statistics
2) Asymptotic analysis

1. Formal definition of the game

$$\mathcal{S}_D = \{\Lambda_0 : P_{fp} \le 2^{-n\lambda}\}$$

$$\mathcal{S}_A = \{z^n : d(y^n, z^n) \le D_{max}\} = \{\underbrace{S_{YZ} : S_Y = P_{y^n}, \textstyle\sum_{ij} S_{YZ}(i,j) d(i,j) \le D_{max}}_{\mathcal{A}_n(D_{max}, P_{y^n}) \text{ set of the admissible maps}}\}$$

transportation map

$$u(\Lambda_0, f) = u_D = -P_{fn}$$

2. Solution of the game

✓ Optimum strategies: S_D^* , S_A^*

$$\Lambda_0^* = \{P \in P_n : \mathcal{D}(P_{x^n}||P_X) < \lambda - |\mathcal{X}|\frac{\log(n+1)}{n}\}$$

$$S_{YZ}^*(i,j;y^n) = \arg\min_{S_{YZ} \in \mathcal{A}(D_{max}, P_{y^n})} \mathcal{D}(S_Z||P_X)$$

✓ The profile (Λ_0^*, S_{YZ}^*) is the *only rationalizable equilibrium* for the game;

✓ The value of the payoff at the equilibrium u^* and the error exponent $\varepsilon = \varepsilon(\lambda)$.

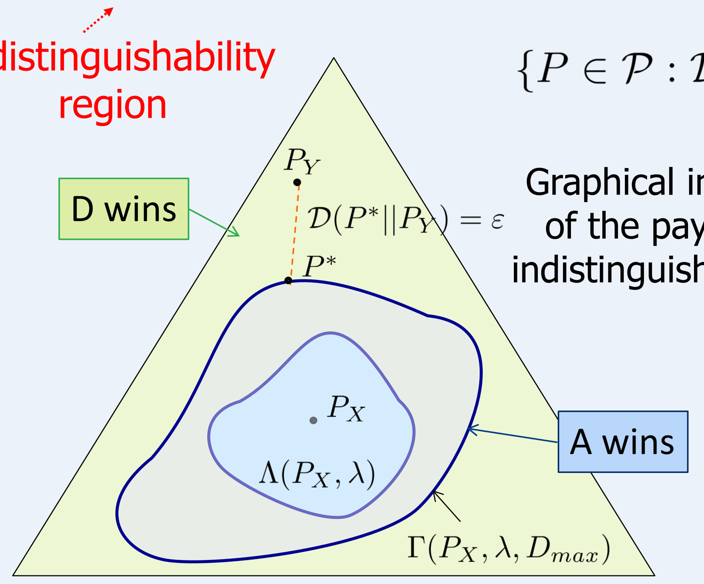
$$\begin{cases} \text{if } P_Y \in \Gamma(P_X, \lambda, D_{max}), & \text{then } P_{fn} \to 1 \ (\varepsilon = 0) \\ \text{if } P_Y \notin \Gamma(P_X, \lambda, D_{max}), & \text{then } P_{fn} \to 0 \ (\varepsilon = \min_{P\in\Gamma} \mathcal{D}(P||P_Y)) \end{cases}$$

→ D wins

→ A wins

$$\Gamma(P_X, \lambda, D_{max}) = \{P \in \mathcal{P} : \exists\, Q \in \underbrace{\Lambda_0(P_X, \lambda)}_{\{P \in \mathcal{P} : \mathcal{D}(P||P_X) \le \lambda\}} \text{ s.t } EMD(P,Q) \le D_{max}\},$$

Indistinguishability region



Graphical interpretation of the payoff and the indistinguishability region

Earth-Mover Distance

$$\min_{\mathbf{S_{YZ}:S_Y=P, S_Z=Q}} \sum \mathbf{S_{YZ}(i,j) d(i,j)}$$

Optimum transportation problem (OTP): looks for the transportation map that moves P into Q minimizing the overall transportation cost.

[1] M.Barni, B.Tondi, The Source Identification Game: An Information-Theoretic Perspective, IEEE Trans. on Information Forensics and Security, vol.8, no.3, pp 450-463, March 2013.

BEST ACHIEVABLE PERFORMANCE FOR D

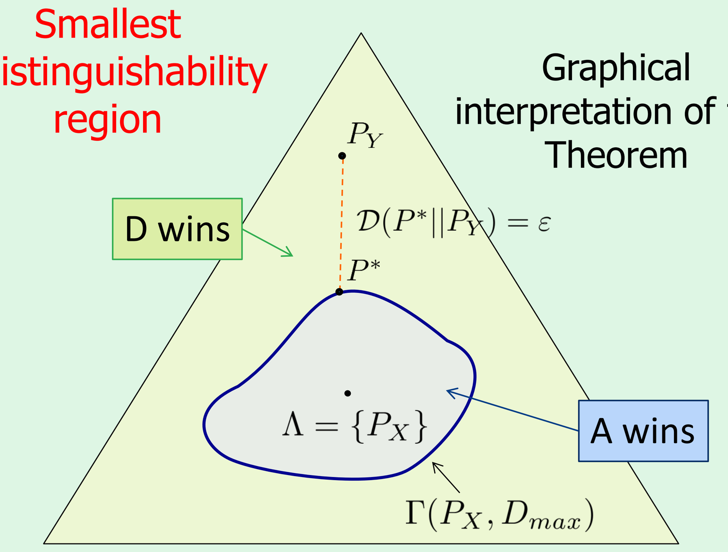
We study the behavior of the indistinguishability region $\Gamma(P_X, \lambda, D_{max})$ when $\lambda \to 0$ (counterpart of Stein's lemma under adversarial conditions).

Theorem: given X ~ P_X and Y ~ P_Y and a maximum allowed per-letter distortion D_{max}, *the maximum achievable false negative error exponent* ε for the SI_{ks} game is

$$\lim_{\lambda\to 0} \lim_{n\to\infty} -\frac{1}{n}\log P_{fn} = \min_{P \in \Gamma(P_X, D_{max})} \mathcal{D}(P||P_Y)$$

where $\Gamma(P_X, D_{max}) = \{P \in \mathcal{P} : \text{EMD}(P, P_X) \le D_{max}\}$

Smallest indistinguishability region



Graphical interpretation of the Theorem

Reminder: the *maximum achievable false negative error exponent* for attack-free hypothesis testing (Stein's lemma) is $\mathcal{D}(P_X||P_Y)$

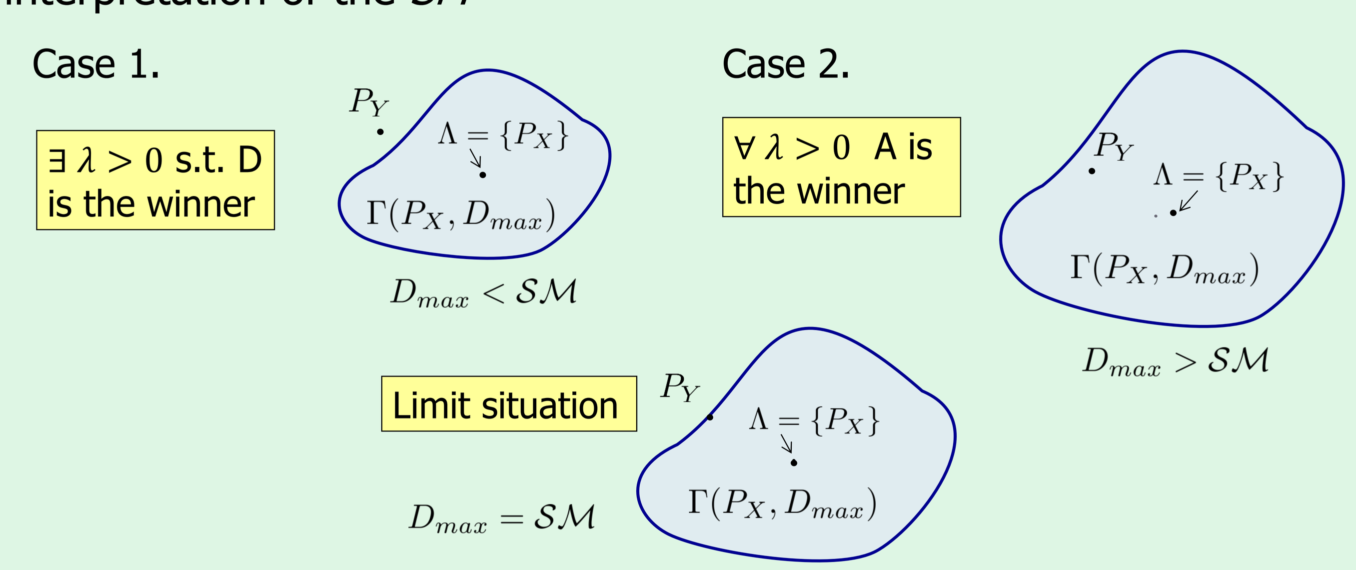
Remark: when $D_{max} = 0$ (no distortion allowed to the attacker) the analysis boils down to the attack-free setup

THE SECURITY MARGIN

Definition (Security Margin). Let X ~ P_X and Y ~ P_Y be two discrete memoryless sources. The *maximum distortion* for which the two sources can be reliably distinguished in the SI_{ks} setting is called Security Margin (*SM*) and is given by

$$\mathcal{SM}(P_Y, P_X) = \text{EMD}(P_Y, P_X)$$

Evaluating the ultimate performances of the game: a graphical interpretation of the *SM*



Properties:

- The *SM* is a symmetric function
- The *SM* is a metric when the distortion measure $d(\cdot)$ adopted by A is a metric

Remark: the *SM* can be computed numerically, by means of the several efficient algorithms already available for the computation of the EMD.

SECURITY MARGIN COMPUTATION

There are some simple cases in which the *SM* can be computed by resorting to analytical computations.

NOTABLE EXAMPLES

- Bernoulli sources: $P_X(1) = p,\ P_Y(1) = q \longrightarrow \mathcal{SM}(P_X, P_Y) = |p - q|$
- Continuous sources:

We adopt the squared Euclidean norm L_2^2 (i.e. distortion function $d(i,j) = |i - j|^2$)

The EMD can be interpreted as the squared *Mallow distance*, then

$$\mathcal{SM}_{L_2^2}(P_X, P_Y) = \min_{P_{XY}: \sum_x P_{XY} = P_Y,\ \sum_y P_{XY} = P_X} E_{XY}[(X-Y)^2]$$

From the decomposition theorem:

$$E_{XY}[(X-Y)^2] = \underbrace{(\mu_X - \mu_Y)^2}_{\text{difference in location}} + \underbrace{(\sigma_X - \sigma_Y)^2 + 2[\sigma_X\sigma_Y - covXY]}_{\text{.... in spread}}$$

difference in **location** .... in **spread**

Then, in order to find the *SM* we have to compute:

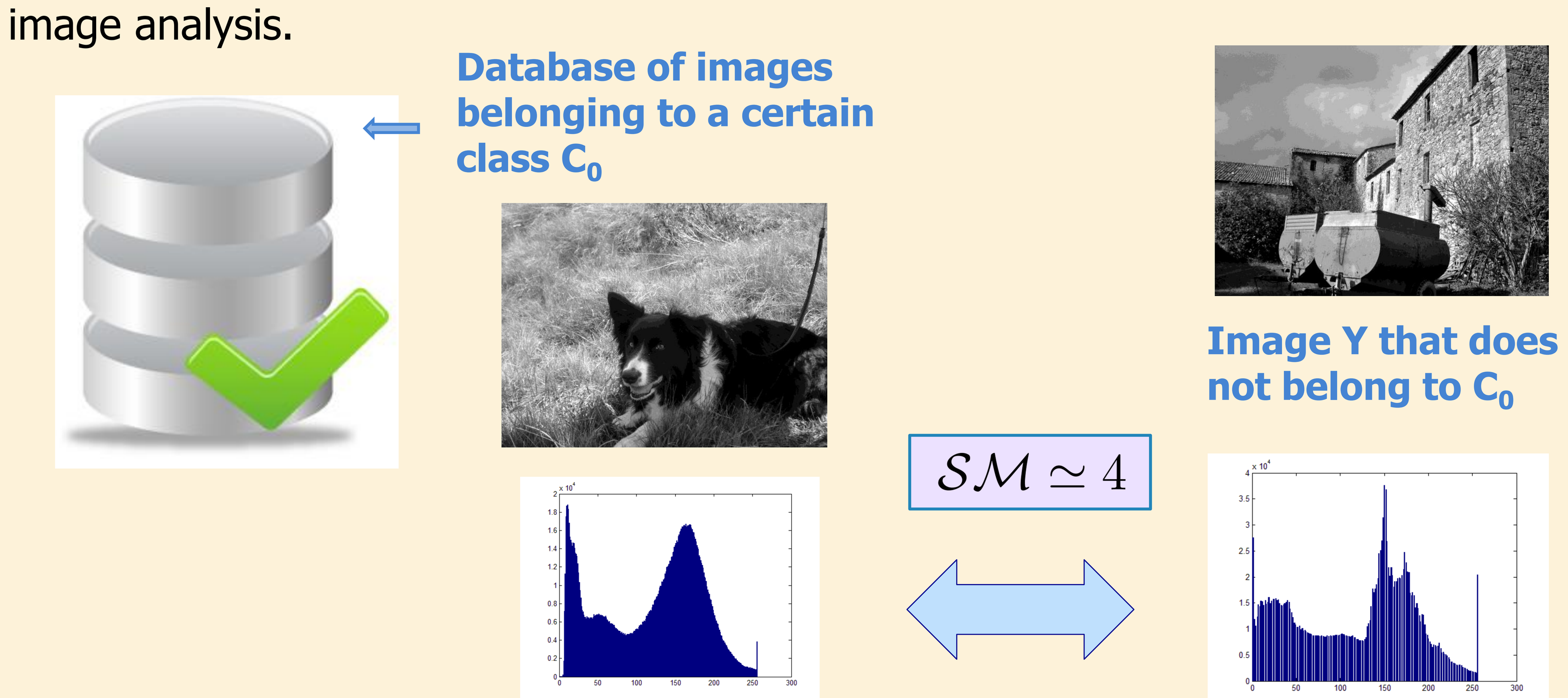
➢ General upper bound for *SM* : $\mathcal{SM}_{L_2^2}(P_X, P_Y) \le (\mu_X - \mu_Y)^2 + \sigma_X{}^2 + \sigma_Y{}^2$

➢ **Remark**: when X and Y belongs to the same class (e.g. Gaussian, Laplacian,…), the *SM* takes a simple and interesting expression in which the shape term vanishes:

$$\mathcal{SM}_{L_2^2}(P_X, P_Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$$

A PRACTICAL MEANING OF THE SECURITY MARGIN

Inspired by the forensic application in [2], we consider an application to histogram-based image analysis.



The *minimum SM* between the histogram of Y and those of the images in the database gives the minimum effort required to the attacker to make Y indistinguishable from the images in C_0

[2] M.Barni, M.Fontani, B.Tondi, A Universal Attack Against Histogram-Based Image Forensics, International Journal of Digital Crime and Forensics (IJDCF) 5 (3), 18